Visibility-Aware Pixelwise View Selection for Multi-View Stereo Matching

Minglun Gong

School of CS, University of Guelph

(Based on the M.Sc. thesis of Zhentao Huang)

IMPROVE LIFE.

Outline

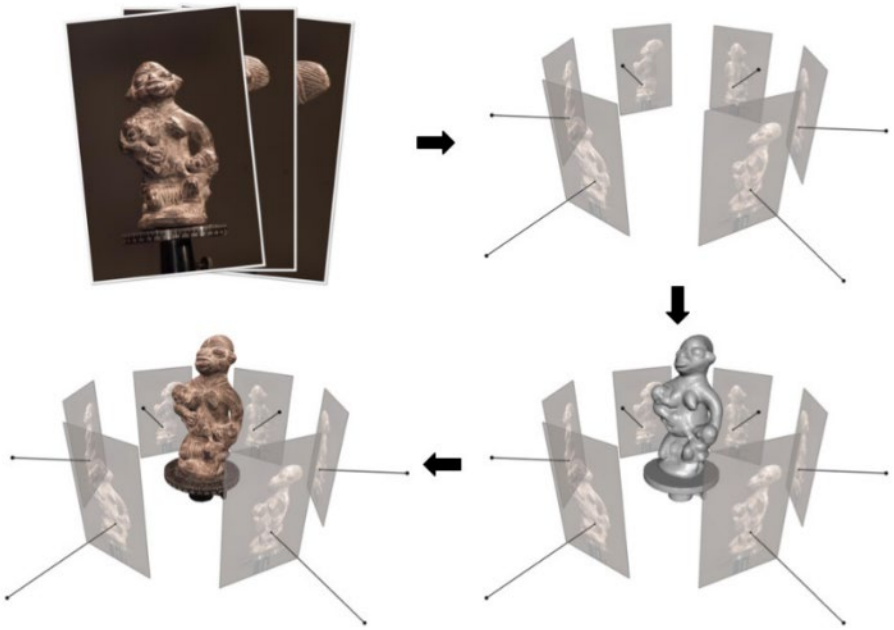
- Introduction of Multi-View Stereo
- Related Works
- Algorithm
  - Pixelwise View Selection
  - Artificial Multi-Bee Colony Algorithm
  - Geometric Consistency Check
- Evaluation Results
- Conclusion and Future Work

What is Multi-View Stereo

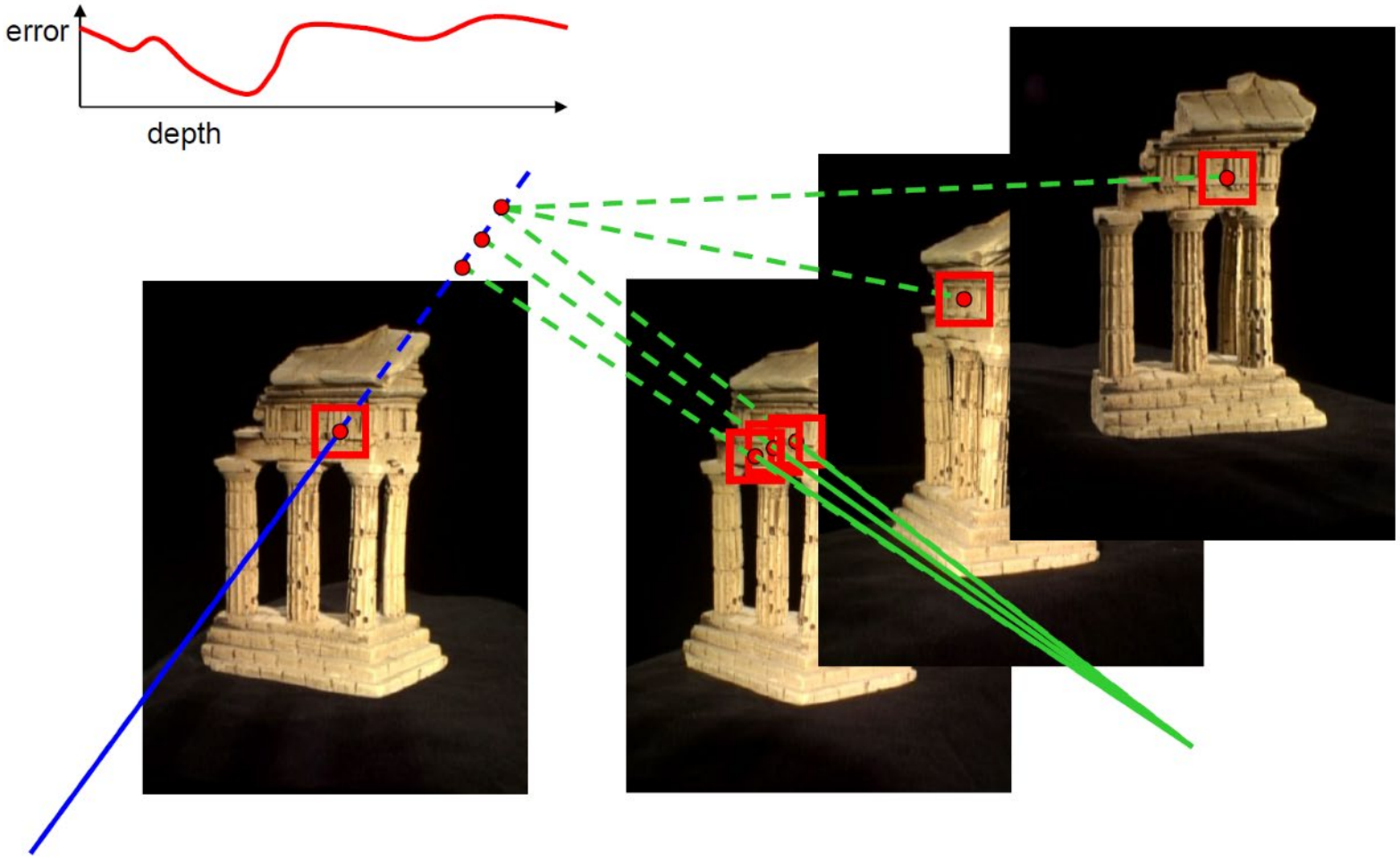
- Taking multiple 2D images from different views and convert them into 3D models.
  - “Given a set of photographs of an object or a scene, estimate the most likely 3D shape that explains those photographs, under the assumptions of known materials, viewpoints, and lighting conditions.”
- Applications:
  - 3D reconstruction of environments or objects for VR/AR
  - Digitization of archeological sites, sculptures.
  - Smart cities

Generic MVS Pipeline

- Collect images from different viewpoints
- Compute camera parameters for each image
- **Reconstruct the 3D geometry of the scene from the set of images and corresponding camera parameters**
- Reconstruct the materials of the scene (optional)



Basic Idea



Related Works: Conventional

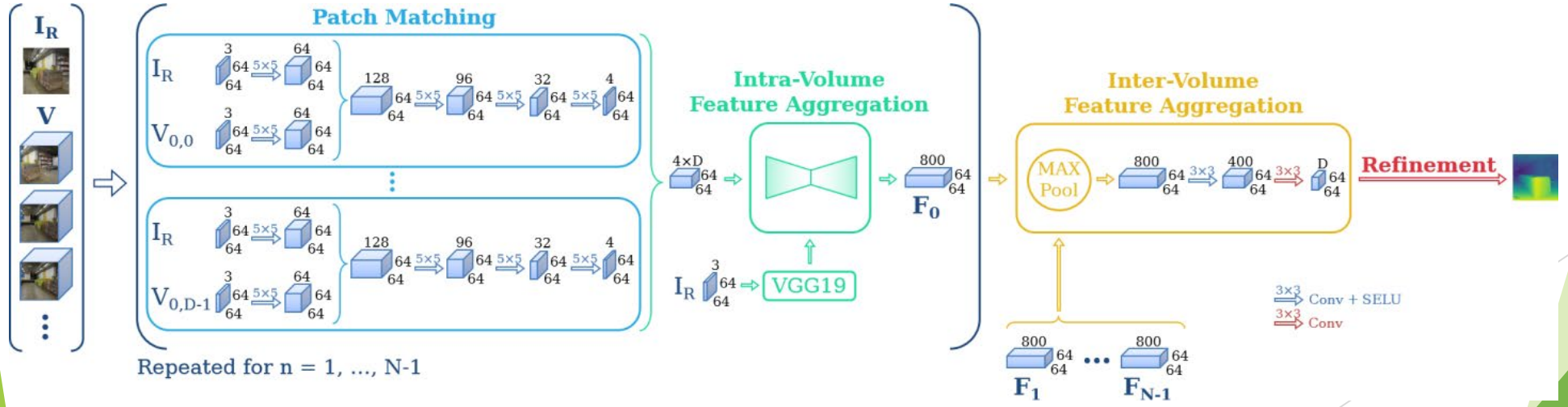
- Gipuma (ICCV 2015)
  - Build on the Patchmatch idea
  - Start from randomly generated 3D planes, the best-fitting planes are iteratively propagated & refined to obtain a 3D depth and normal field per view.
- COLMAP (ECCV 2016)
  - Jointly estimate pixel-wise view selection, depth map, & surface normal.
- ACMH (CVPR 2019)
  - Adopt adaptive checkerboard sampling, multi-hypothesis joint view selection, & multi-scale geometric consistency guidance.
- ACMMP (TPAMI 2023)
  - Extension on ACMH
  - Combine the multi-scale geometric consistency with planar priors.

Related Works: Learning-based

- Volumetric-based approaches:
  - SurfaceNet (ICCV 2017)
  - LSM (NIPS 2017)
- Predict volume occupancy in 3D space.
  - Use 3D neural networks to predict the on-surface probability for each voxel position.
  - Working on 3D volume has high memory requirement, which limits the reconstruction resolution.
- Depth-map-based methods:
  - DeepMVS (CVPR 2018)
  - MVSNet (ECCV 2018)
  - PatchmatchNet (CVPR 2021)
- Predict a depth map for each view.
  - Use convolution networks to perform local cost aggregation, which determine the depth of each pixel.
  - Additional steps needed to fuse depth maps into 3D model.

DeepMVS (CVPR 2018)

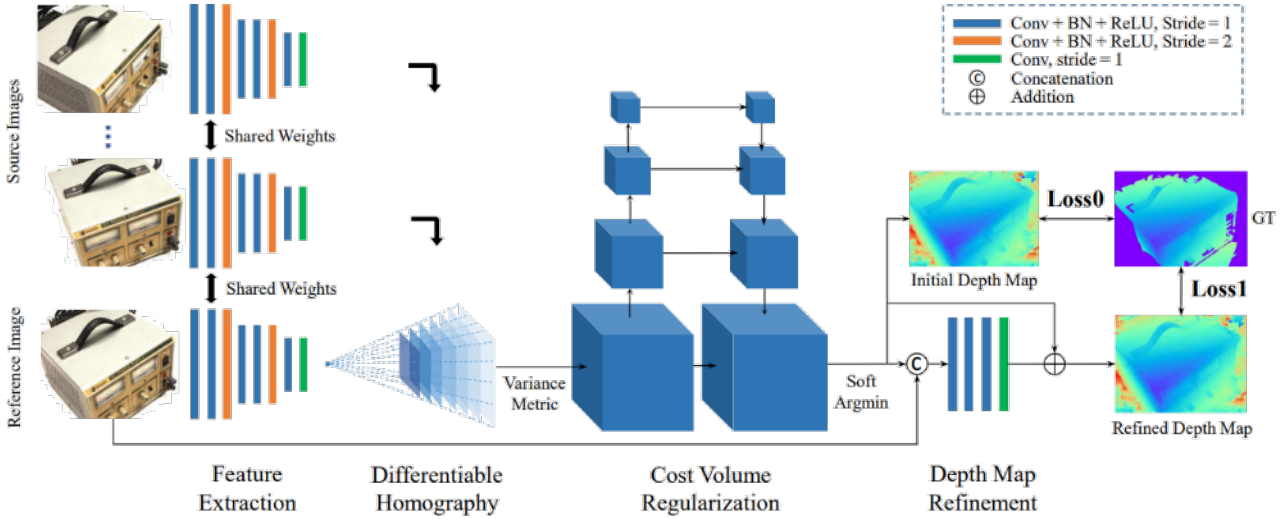
- Take an arbitrary number of posed images as input and produce a set of plane-sweep volumes.
  - Contain a patch matching network, an intra-volume, & an inter-volume feature aggregation network.
  - Integrate multi-layer feature activations from a VGG-19 network pretrained on a photorealistic synthetic dataset.



UNIVERSITY of GUELPH
IMPROVE LIFE.

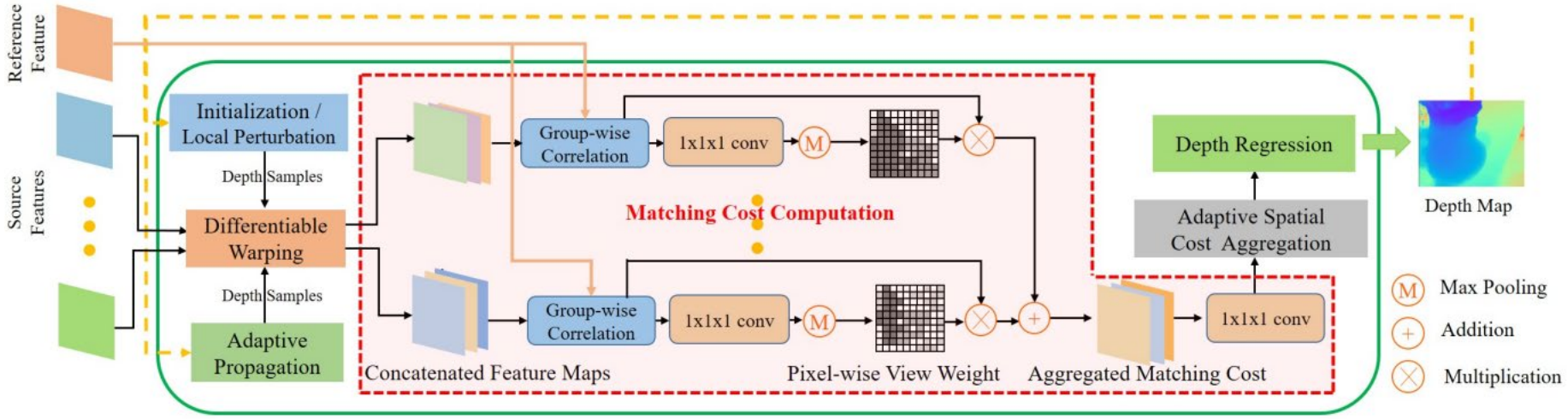
MVSNet (ECCV 2018)

- Build 3D cost volume in the camera frustum instead of the Euclidean space.
  - Use a variance-based metric to map multiple features into one cost feature in the volume.
  - Treat all source views equally using a variance-based cost metric.
    - *Does not handle occlusion*



PatchmatchNet (CVPR 2021)

- Extend the original Patchmatch idea into the deep learning era
  - Decrease memory consumption & run-time for high-resolution MVS.
    - *Embed the end-to-end trainable model into a coarse-to-fine framework to speed up computation.*
  - Augment the traditional propagation & cost evaluation steps with learnable, adaptive modules



Limitations of Existing Work

Conventional Approaches

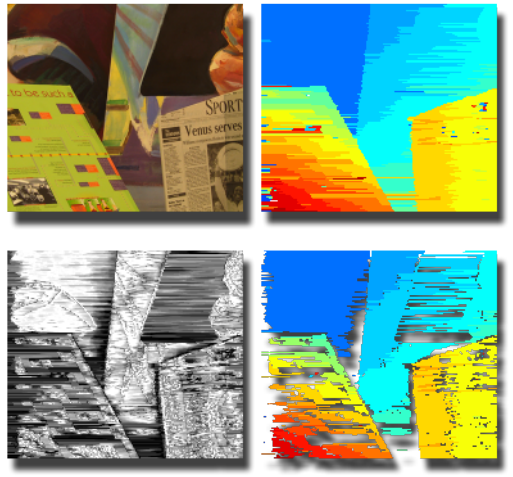
- Use ad hoc approach to handle occlusions.
  - For example, select the top k matches that have the lowest costs.
  - Do not consider geometric relations.
- Output noisy results in weakly textured regions.

Learning-based Approaches

- The quality of the reconstruction is limited by the diversity of training dataset (Wang et al., 2023).
- Volume-based approaches use 3D CNN, which are time & memory consuming.
- Most require a fixed-size of image input and assume front-parallel surfaces.

Motivation

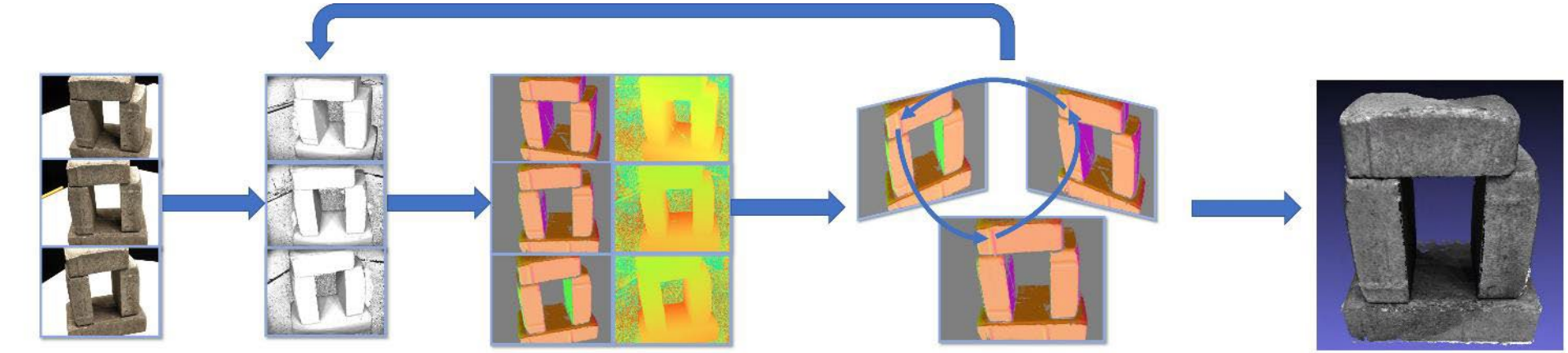
- Introduce explicit occlusion detection & handling to MVS.
  - Extend my previous work on occlusion modeling for binocular stereo.
- Apply an efficient optimization technique.
  - Use artificial multi-bee-colony algorithm to search and propagate optimal solutions.



TPAMI 2005

GECCO 2016

Algorithm Overview

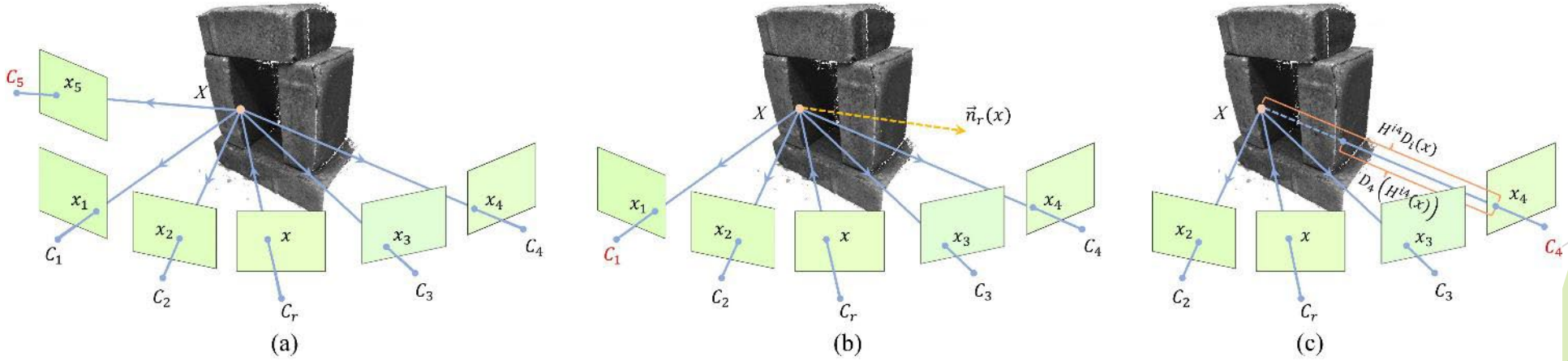


Reference View Selection

- When reconstructing large-scale scenes, the total number of views needed is in hundreds.
  - Using all other views as source views when calculating for a reference view, as in MVSNet, is both computational expensive and problematic.
- Source view selection strongly impacts the quality of the reconstruction result.
  - For a given pixel in a reference view, the corresponding pixels in a source view may not be visible.
  - Which set of source views should be used is pixel dependent.

Our Visibility-aware Approach

- Pixelwise & progressive selection based on available information.
  - First use a triangulation term to remove source views with large view angle difference.
  - Once the normal for x is estimated, source views with poor incident angles will be removed.
  - Finally, validated solutions are used to further remove occluded views from the set.



UNIVERSITY of GUELPH
IMPROVE LIFE.

Artificial Multi-Bee Colony Algorithm

- The search space for MVS is huge.
  - Find depth & normal (4D) for each pixel in each reference image.
- AMBC is initially proposed for solving the k-nearest-neighbor fields (k-NNF) problem.
  - Represent each solution (depth & normal) as a food source.
    - *Evaluate the fitness of the food source using bilaterally weighted Normalized Cross-Correlation.*
  - Hold 10 food sources for every colony (pixel).
  - Send out three kinds of bees: Employed bees, Onlooker bees, Scout bees.
    - *When a better food source is found, replace the worst one.*

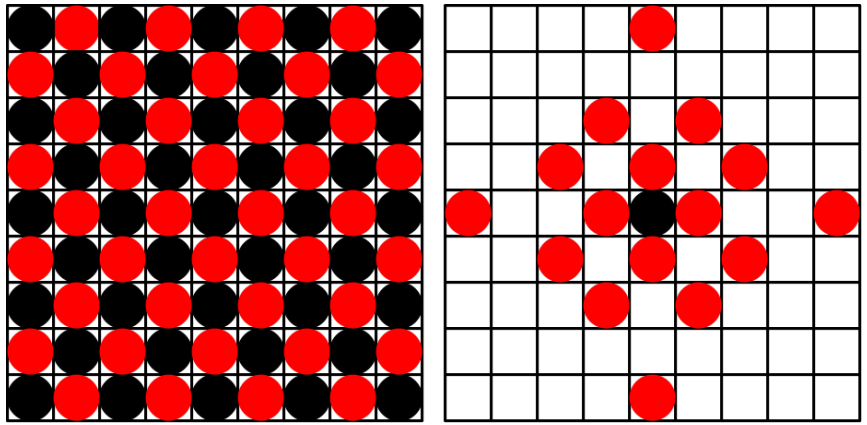
Employed Bees (Local Search)

- Perform search within the local colonies.
  - The new food source is generated by perturbing the existing food sources.
  - If the new food source has lower matching cost, replace the previous food source.

$$y'_x = y_x + R(-1, 1)(y_n - y_x)$$

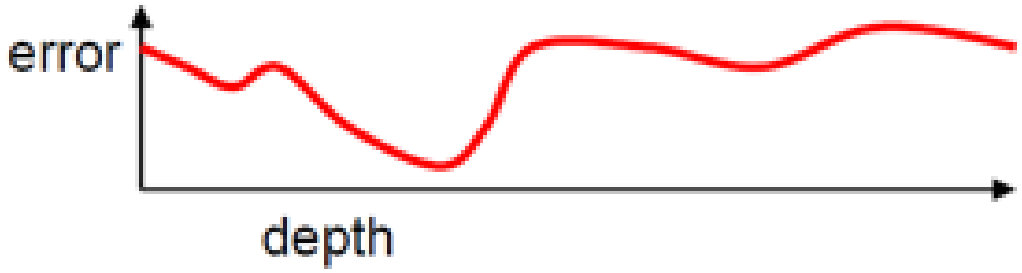
Onlooker Bees (Spatial Propagation)

- Propagate good solutions to neighboring colonies.
  - Divide the image into red and black groups.
  - Pixels in the same group can be processed in parallel without interfering with others.
  - If the food source from a neighbor colony has lower matching cost, use it to replace the worst food source in the current colony.



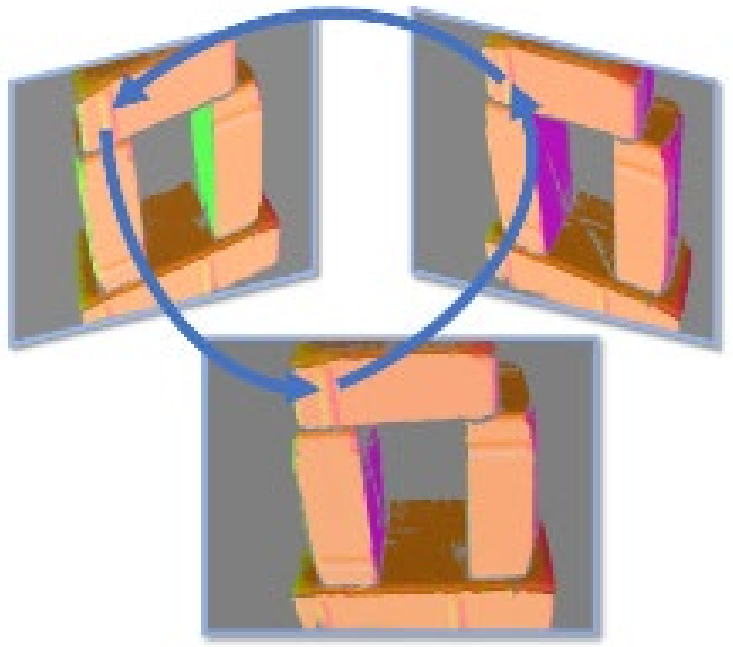
Scout Bees (Avoid Local Optimal)

- Employed bees & Onlooker bees only perturb or propagate existing solutions.
  - Food sources will converge to a relatively small range after several iterations.
- Scott bees perform global searching and avoid local optimal.
  - Every time the Employed/Onlooker bees do not update the food source, it will be counted as one trial.
  - For the top-2 to worst, if their trial count is larger than 10, replace it with a randomly generated food source.



Geometry Consistency Check

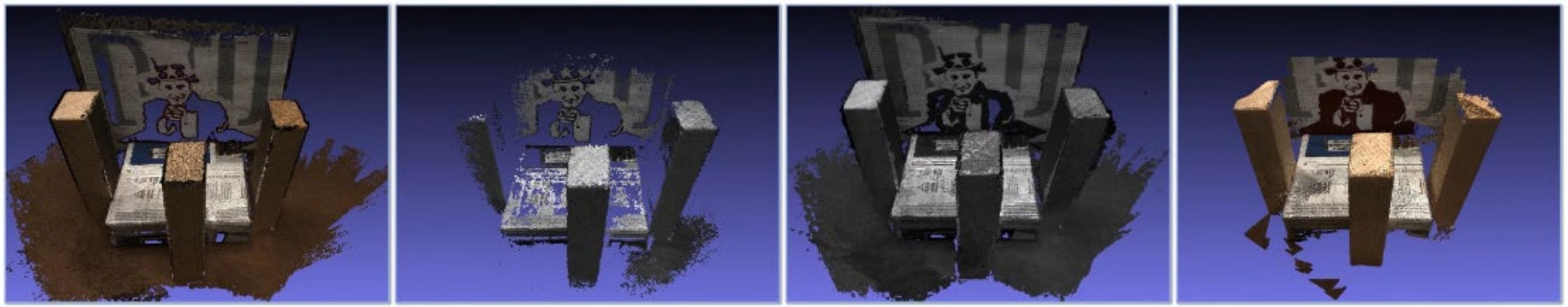
- Filter out mismatches though comparing results obtained for different views.
  - Pixel-wise visibility information is used again.
- Verified solutions will be:
  - Propagated to those unverified pixels in other selected views (Propagation Between Views).
  - Added a reward in the cost during the Onlooker Bees searching (Smoothness Constraint).



Datasets

- DTU Robot Image dataset (main)
  - Contains 124 scenes
  - 49 or 64 images with 1600x1200 resolution.
  - Captures objects at a close distance and hence visibility handling is a major concern.
- Tanks and Temples dataset
  - Contains 21 scenes divided into training and test sets.
  - Test sets are further split into “Intermediate” & “Advanced” subsets.
  - Videos with 1920x1080 resolution.
- There are other widely-used MVS dataset such as ETH3D, BlendedMVS…

Visual Comparison on DTU dataset

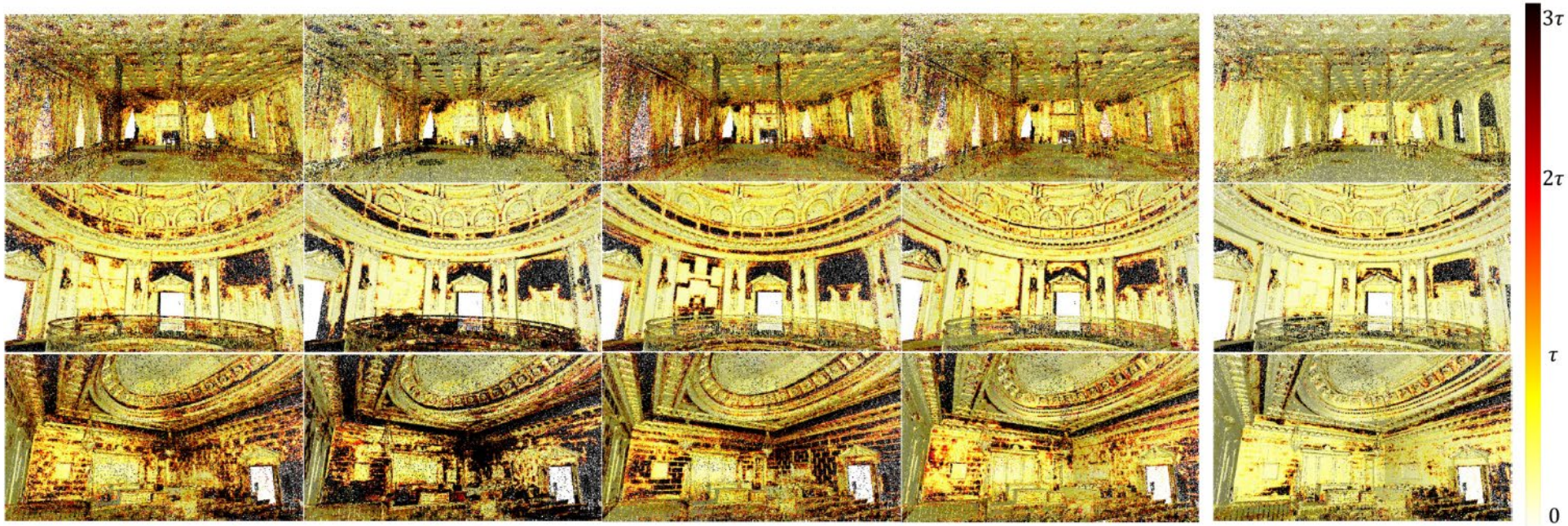


(a) COLMAP (b) Gipuma (c) Ours (d) Ground Truth

Quantitative Evaluation on DTU dataset

Conventional	Accuracy (mm)	Completeness (mm)	Overall
Furukawa	0.613	0.941	0.777
Tola	0.342	1.190	0.766
Campbell	0.835	0.554	0.695
Gipuma	0.283	0.873	0.578
COLMAP	0.411	0.657	0.534
Ours	0.405	0.381	0.393
Learning-based	**Accuracy (mm)**	**Completeness (mm)**	**Overall**
SurfaceNet	0.450	1.040	0.745
MVSNet	0.396	0.527	0.462
P-MVSNet	0.406	0.434	0.420
R-MVSNet	0.383	0.452	0.417
PatchMatchNet	0.427	0.277	0.352
UniMVSNet	0.352	0.278	0.315

Visual Comparison on Tanks & Temples



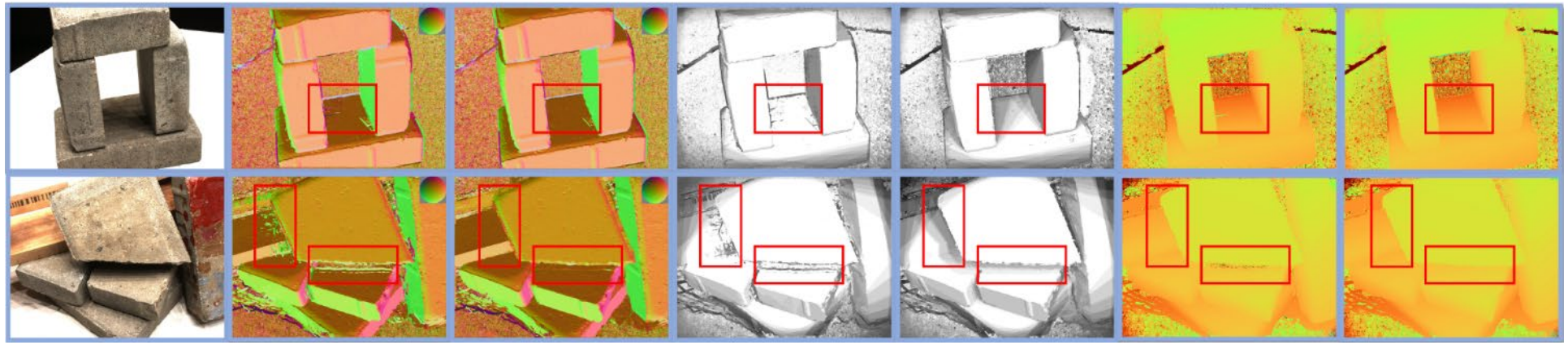
(a) EPP-MVSNet (b) CasMVSNet (c) ACMH (d) ACMP (e) Ours

Quantitative Evaluation on Tanks & Temples

	Method	Intermediate									Advanced						
		mean	Fam.	Fra.	Hor.	Lig.	M60	Pan.	Pla.	Tra.	**mean**	Aud.	Bal.	Cou.	Mus.	Pal.	Tem.
Non-L.-based	COLMAP[40]	42.14	50.41	22.25	26.63	56.43	44.83	46.97	48.53	42.04	27.24	16.02	25.23	34.70	41.51	18.05	27.94
	PLC[27]	54.56	70.09	50.30	41.94	59.04	49.19	55.53	56.41	54.13	34.44	23.02	30.95	42.50	49.61	26.09	34.46
	ACMH[57, 58]	54.82	69.99	49.45	45.12	59.04	52.64	52.37	58.34	51.61	33.73	21.69	32.56	40.62	47.27	24.04	36.17
	ACMM[58]	57.27	69.24	51.45	46.97	63.20	55.07	57.64	60.08	54.48	34.02	23.41	32.91	41.17	48.13	23.87	34.60
	ACMP[59]	58.41	70.30	54.06	**54.11**	61.65	54.16	57.60	58.12	57.25	37.44	**30.12**	34.68	**44.58**	50.64	27.20	37.43
	ACMMP[57]	59.38	70.93	55.39	51.80	63.83	55.94	59.47	59.51	**58.20**	37.84	30.05	35.36	44.51	50.95	27.43	38.73
	Ours	54.53	68.09	53.46	40.67	58.28	54.91	53.79	54.10	52.93	**38.26**	24.97	**44.25**	41.57	**53.11**	28.52	37.11
Learning-based	MVSNet[62]	43.48	55.99	28.55	25.07	50.79	53.96	50.86	47.90	34.69	-	-	-	-	-	-	-
	PatchMatchNet[52]	53.15	66.99	52.64	43.24	54.87	52.87	49.54	54.21	50.81	32.31	23.69	37.73	30.04	41.80	28.31	32.29
	CVP-MVSNet[61]	54.03	76.5	47.74	36.34	55.12	57.28	54.28	57.43	47.54	-	-	-	-	-	-	-
	UCS-Net[7]	54.83	76.09	53.16	43.03	54.00	55.60	51.49	57.38	47.89	-	-	-	-	-	-	-
	CasMVSNet[15]	56.84	76.37	58.45	46.26	55.81	56.11	54.06	58.18	49.51	31.12	19.81	38.46	29.10	43.87	27.36	28.11
	AttMVS[30]	60.05	73.90	62.58	44.08	**64.88**	56.08	59.39	**63.42**	56.06	31.93	15.96	27.71	37.99	52.01	29.07	28.84
	GBi-Net[36]	61.42	**79.77**	**67.69**	51.81	61.25	60.37	55.87	60.67	53.89	37.32	29.77	42.12	36.30	47.69	**31.11**	36.93
	EPP-MVSNet[31]	**61.68**	77.86	60.54	52.96	62.33	**61.69**	**60.34**	62.44	55.30	35.72	21.28	39.74	35.34	49.21	30.00	**38.75**

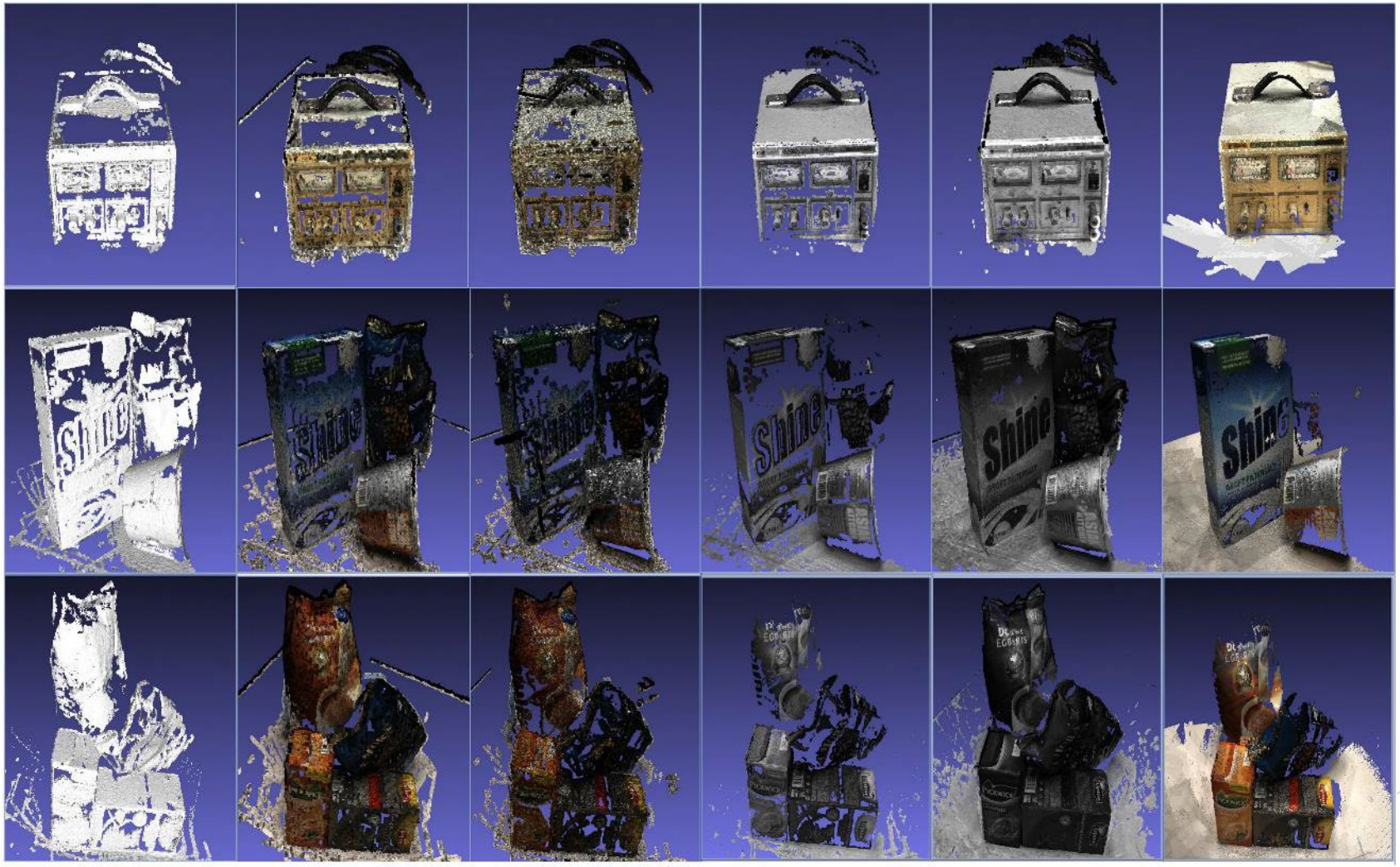
Ablation Study on Pixelwise View Selection

Input | Normal Map | Source View Set Map | Depth Map



Baseline | Baseline+PVS | Baseline | Baseline+PVS | Baseline | Baseline+PVS

Ablation Study on Smoothness Constraint

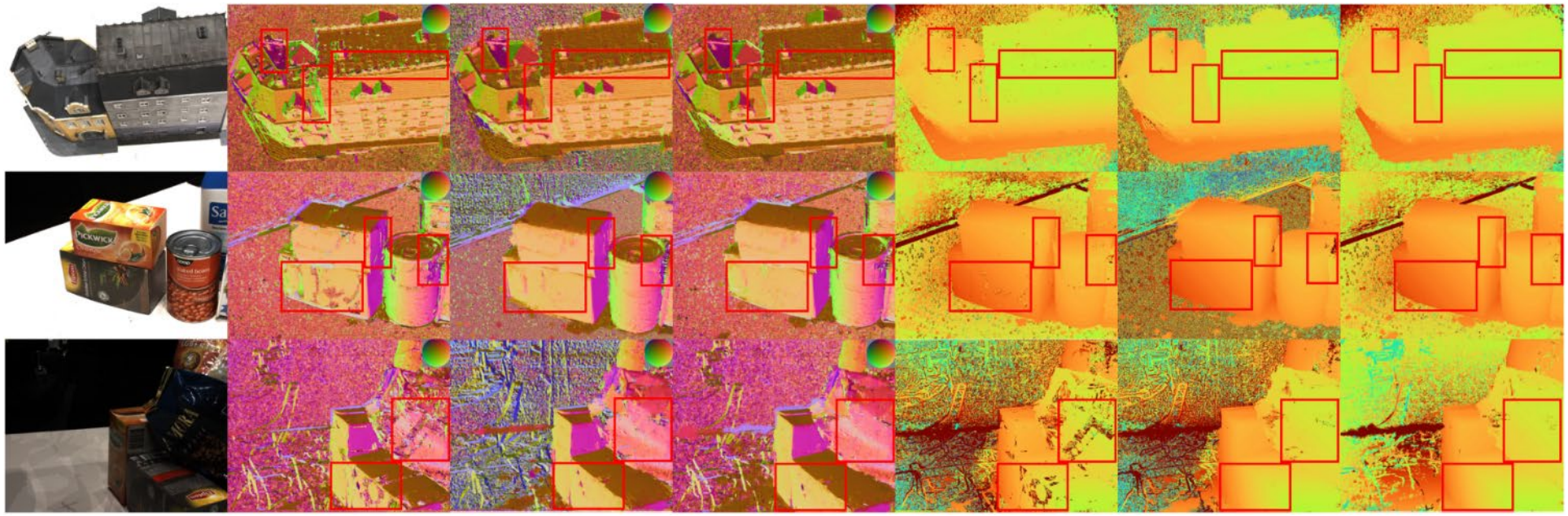


(a) Tola (b) Furukawa (c) Camp (d) Ours w/o SC (e)Ours (f) Ground Truth

UNIVERSITY of GUELPH

IMPROVE LIFE.

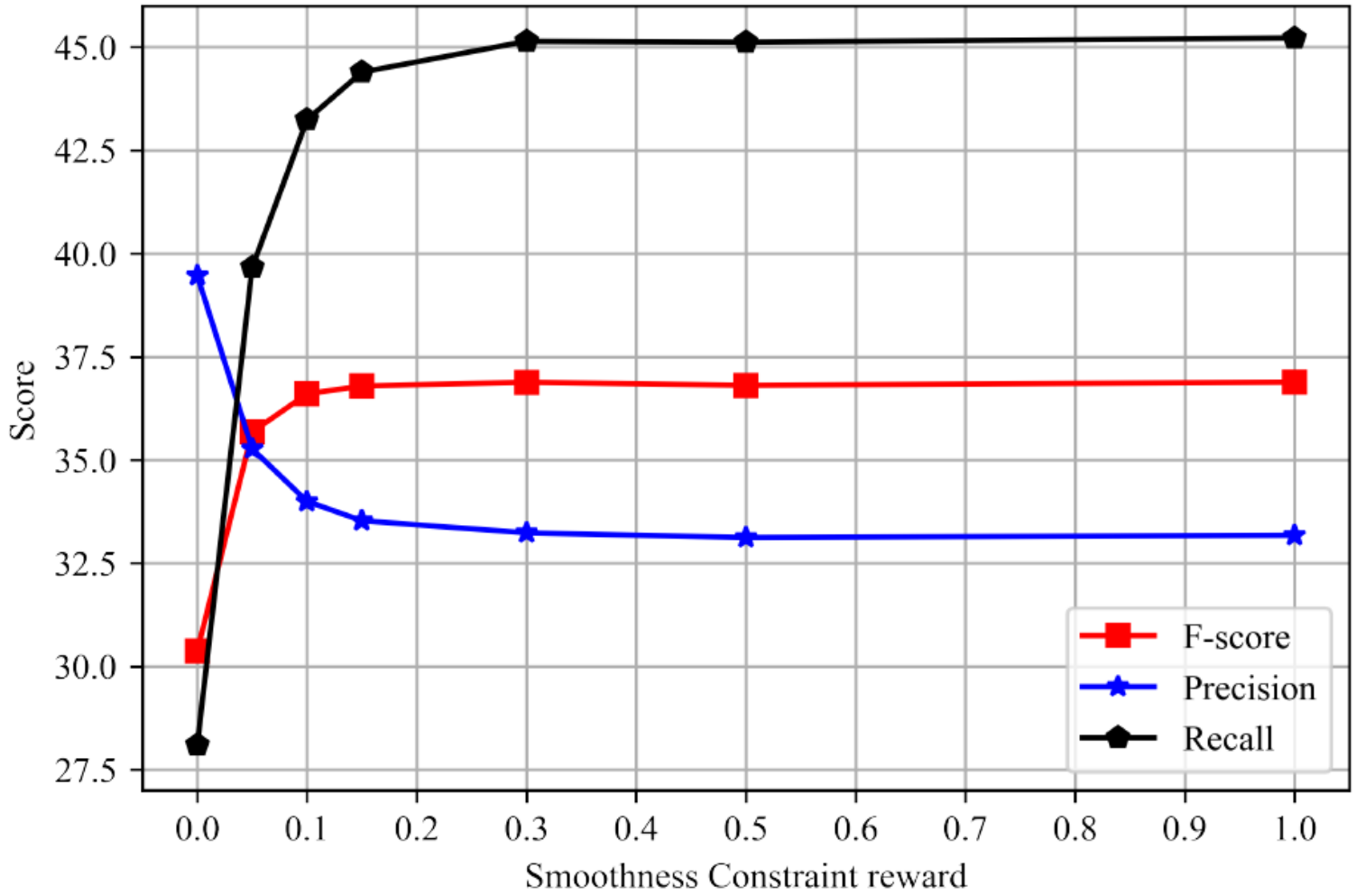
Ablation Study on Both (Visually)



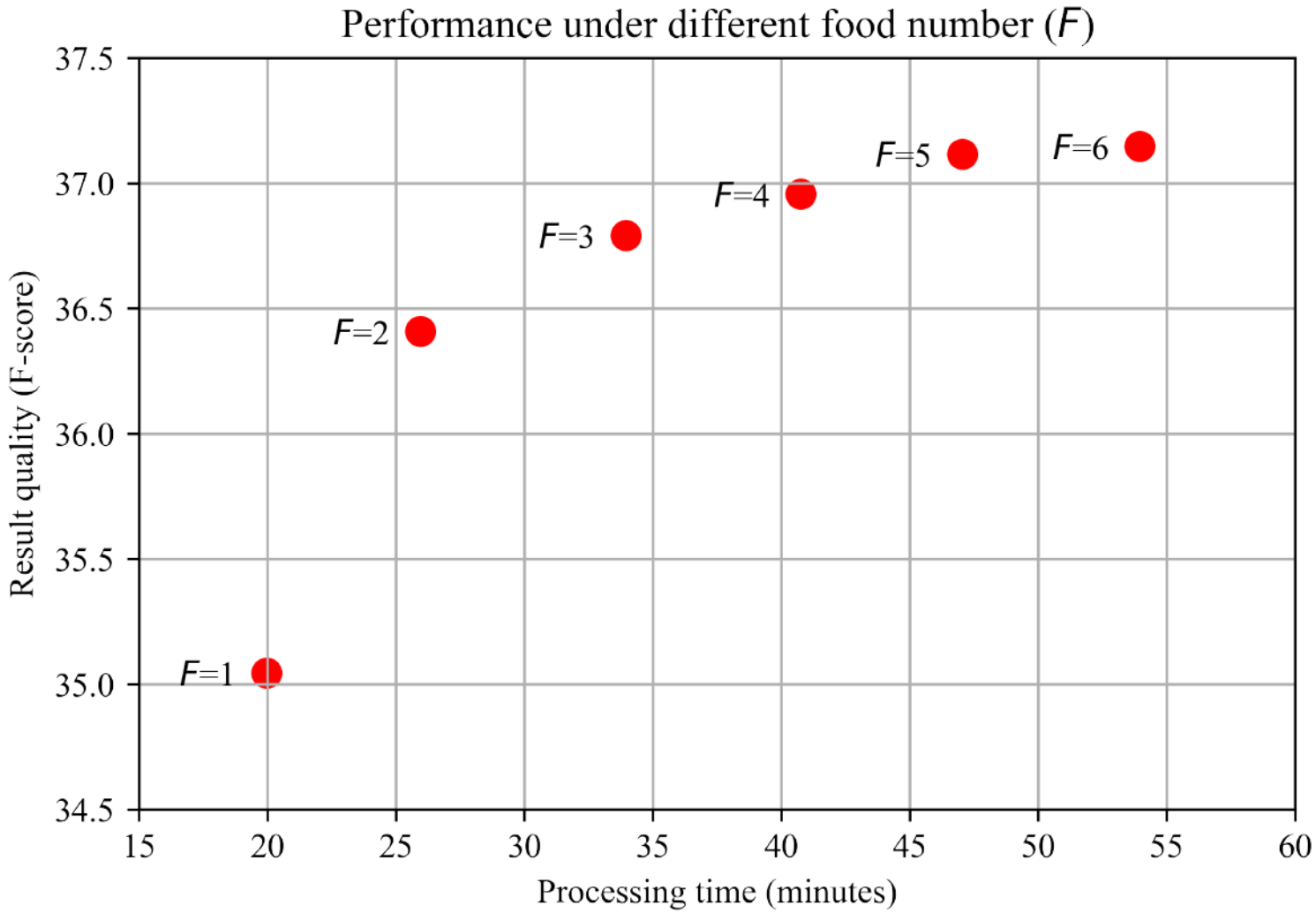
Ablation Study on Both (Quantitatively)

Method	Artificial Multi-Bee Colony			SC	PVS	Acc.	Comp.	Overall
	Employed.	Onlooker.	Scout.					
Baseline	✓	✓	✓			0.363	0.501	0.432
Baseline + SC	✓	✓	✓	✓		0.408	0.409	0.409
Baseline + PVS	✓	✓	✓		✓	**0.362**	0.453	0.407
Baseline + SC + PVS (**Ours**)	✓	✓	✓	✓	✓	0.405	**0.381**	**0.393**

Ablation Study on Smoothness Parameter



Ablation Study on Food Number



Conclusion

- Propose a visibility-aware pixelwise view selection method for PatchMatch-based MVS.
  - View selection is progressively improved for individual pixels.
  - Selected source views are used for both matching cost evaluation and consistency check.
- AMBC algorithm is applied to search for optimal solutions.
  - The onlooker bee is used for both inter-image and intra-image solution propagation.
  - Smoothness constraints are applied on verified solutions to tackle the low-textured regions.
- Two datasets are used for the evaluation of the proposed method.
  - Ablation studies demonstrated the effectiveness of each main component.

Future Work

- How to combine learning-based modules with geometry-based constraints?
  - DNN is good at predict depth based on learned priori knowledge, but can be ineffective in detecting occlusions, predicting precise depth, & performing consistency check.
  - Use network to predict depth and/or evaluate matching costs but conventional approach for refinement & consistency check.
- How to handle both geometry details and weakly textured areas?
  - The former requires fine scale, whereas the later coarse scale.
  - Apply scale-invariant learning technique.
- How to handle high-resolution inputs under limited memory?
  - Break the input image into patches and integrate solutions for individual patches together?

QUESTIONS?

Abstract

- The performance of PatchMatch-based multi-view stereo algorithms is greatly influenced by the chosen source views used for matching cost computation. Existing methods usually detect occlusions through rather ad-hoc approaches, which can negatively impact the computation. This talk introduces an innovative approach that deliberately models view visibility. We present a novel visibility-guided pixelwise view selection scheme that progressively refines the set of source views for each pixel in the reference view using visibility information from validated solutions. Furthermore, the Artificial Multi-Bee Colony (AMBC) algorithm is leveraged to parallelly search optimal solutions for different pixels. To ensure smoothness of neighboring pixels and better manage textureless areas, rewards are assigned to solutions that come from validated sources. Our method, validated through experiments on two datasets, improves detail recovery in occluded and low-textured regions, demonstrating state-of-the-art performance on demanding scenes.

Reference

- Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. IJCV, 1-16, 2016.
- Neill DF Campbell, George Vogiatzis, Carlos Hern´andez, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. ECCV, 766-779. Springer, 2008.
- Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. TPAMI, 32(8):1362-1376, 2009.
- Silvano Galliani, Katrin Lasinger, and Konrad Schindler.Massively parallel multiview stereopsis by surface normal diffusion. ICCV, 873-881, 2015.
- Johannes L Sch¨onberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. ECCV, 501-518. Springer, 2016.
- Engin Tola, Vincent Lepetit, and Pascal Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. TPAMI, 32(5):815-830, 2009.

Reference

- Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. ICCV 2307-2315, 2017.
- Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. ECCV, 767-783, 2018.
- Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, and Yawei Luo. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. CVPR, 10452-10461, 2019.
- Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. CVPR, 5525-5534, 2019.
- Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. CVPR, 2495-2504, 2020.